Artificial Intelligence Governance Under Change:
Foundations, Facets, Frameworks

TLDR **— Handout & frameworks for reference**

PhD Defense, Matthijs M. Maas

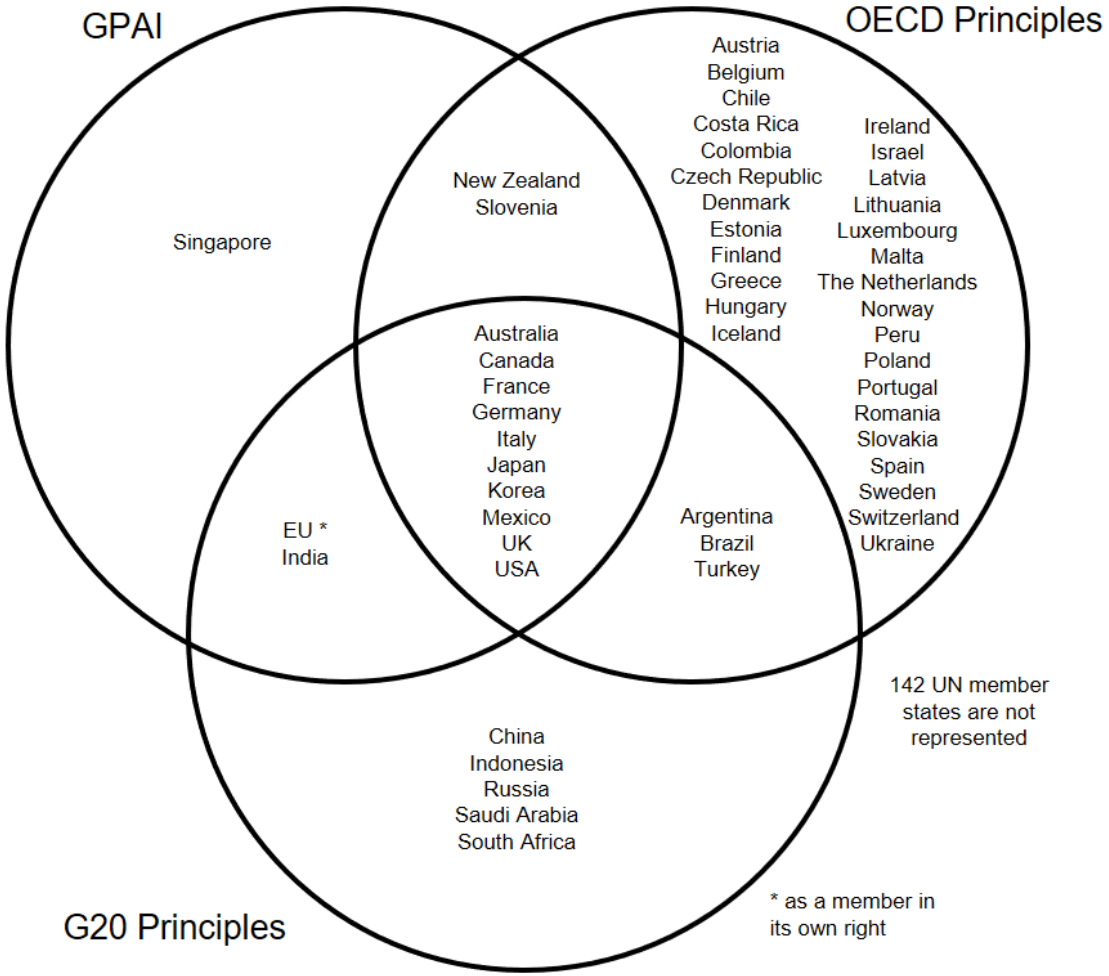
April 21st, 2021

Comments and feedback to Matthijs.Maas@jur.ku.dk

Dissertation – core papers

- **Paper [I]:** Maas, Matthijs M. "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons." *Contemporary Security Policy* 40, no. 3 (February 6, 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464.
- **Paper [II]:** Maas, Matthijs M. "Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot." *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 129–57. https://doi.org/10.1163/18781527-01001006.
- **Paper [III]:** Maas, Matthijs M. "International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order." *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56. https://law.unimelb.edu.au/__data/assets/pdf_file/0005/3144308/Maas.pdf
- **Paper [IV]:** Cihon, Peter, Matthijs M. Maas, and Luke Kemp. "Should Artificial Intelligence Governance Be Centralised? Design Lessons from History." In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, 228-34. New York, NY, USA: ACM, 2020. https://doi.org/10.1145/3375627.3375857.

Fragmented state of AI governance regime complex



Fragmented membership of international AI initiatives
(*as of September 2020: in December 2020, Brazil, the Netherlands, Poland and Spain joined GPAI)

Reproduced (with permission) from:
Cihon, Peter, Matthijs M. Maas, and Luke Kemp. "Fragmentation and the Future: Investigating Architectures for International AI Governance." *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.

Research questions

RQ: ***How should global governance for artificial intelligence account for change?***

A. Why do we require governance strategies for artificial intelligence? **Why do these require new strategies for change?**

B. Why, when, and how should governance systems approach and respond to **AI-driven sociotechnical change**?

C. Why, when, and how might AI applications **disrupt global governance**, by driving or necessitating changes to its substance and norms, its processes and workings, or its political scaffolding?

D. Why and how might **changes in the broader global governance architecture**, as well as amongst individual AI regimes, affect the prospects, development and efficacy of the 'regime complex' for AI?

E. What **insights can these three conceptual frameworks provide** in exploring the prospects and dynamics of the emerging AI governance regime complex?

Overview: Conceptual Lenses on Change in AI Governance

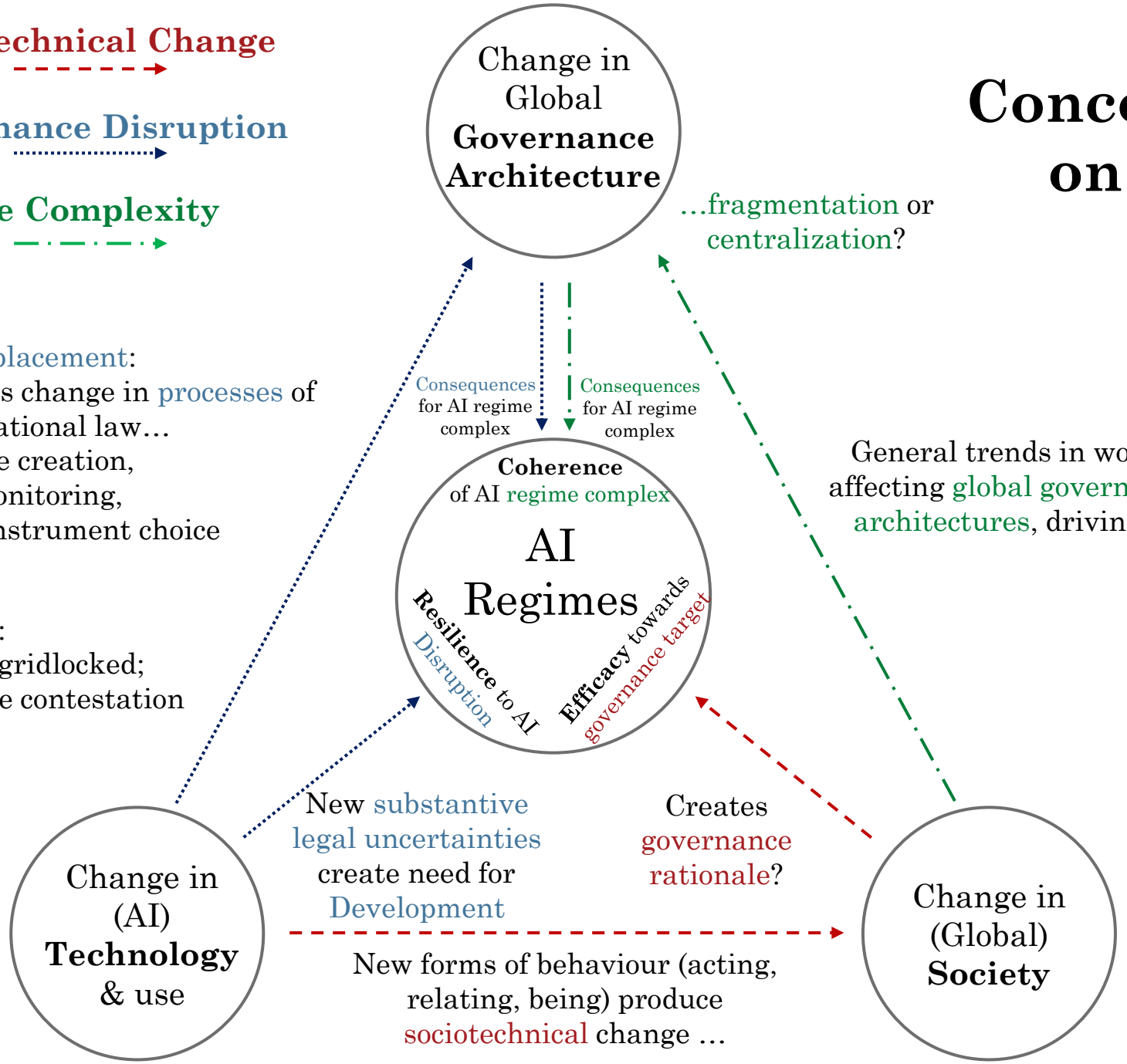
Sociotechnical Change

Governance Disruption

Regime Complexity

Displacement:
Automation drives change in processes of international law…
-rule creation,
-monitoring,
-shifting instrument choice

Destruction:
-'erosion' AI problems gridlocked;
-'decline' AI enables active contestation



UNIVERSITY OF COPENHAGEN

Sociotechnical change: gov. targets & problem logics

Problem Logic and questions	Corresponding governance rationales	Governance Surface (origin / barriers to resolution)	Governance Logics (selected)
Ethical challenges What rights, values or interests does this threaten?	• New risks to moral interests, rights or values • New threats to social solidarity • Threats to democratic process	• Origin: actor **apathy** (to certain values) or **ignorance** • Barriers: underlying societal disagreement (culturally and over time) over how to weigh the values, interests or rights at stake	• Bans ('mend—or end') • Oversight & accountability mechanisms; auditing • 'Machine ethics' • Ethics education • Value-Sensitive Design
Security threats How is this vulnerable to misuse or attack?	• New risks to moral interests, rights or values • New risks to human health or safety	• Origin: Actor **malice** (various motives) • 'Offense-defense balance' of AI knowledge • Barriers: Intrinsic vulnerability of human social institutions to automated social engineering attacks.	• Perpetrator-focused: change norms, prevent access; improve detection & forensics capabilities to ensure attribution and deterrence • Target-focused: reduce exposure; red-teaming; 'security mindset'
Safety risks Can we rely on- and control this?	• New risks to human health or safety	• Origin: Actor **negligence**, automation bias • 'Many hands' problem—long and discrete supply chains • Barriers: Behavioural features of AI systems (opacity; unpredictability; specification gaming)	• Relinquishment (of usage in extreme-risk domains) • 'Meaningful Human Control' (various forms) • Safety engineering (e.g. reliability; corrigibility; interpretability; formal verification etc. etc.) • Liability mechanisms & tort law;
Structural shifts How does this shape our decisions?	• (all, indirectly)	• Origin: Systemic **incentives** for actors (alters choice architectures; increases uncertainty & complexity; competitive value erosion) • Exacerbates other challenges	• Arms control (mutual restraint) • Confidence-Building Measures (increase trust or transparency)
Common Benefits How can we realize opportunities for good with this?	• Possible market failures	• Origin: Systemic **incentives** for actors (Coordination challenges around cost-sharing, free-riding) • Barriers: overcoming loss aversion	• (Global) standards • 'Public interest' regulation and subsidies • 'Windfall clause' & redistributive guarantees
Governance Disruption How does this change how we regulate?	• New risks directly to existing regulatory order	• Origin: Legal system **exposure:** dependence on conceptual orders or assumptions	• Provisions to render governance 'innovation-proof': technological neutrality; authoritative interpreters, sunset clauses; … • Oversight for legal automation; distribution

Governance Disruption

Type			Example
Need for **Development**	**New governance gaps**		• AI-enabled swarm warfare (possibly) not covered by existing international regimes
	Conceptual uncertainty or ambiguity		• LAWS highlight potential ambiguity or inadequacy of concepts such as 'intent', 'effective control', etc.
	Incorrect scope of application (unintentional or engineered)		• Underinclusive application of Convention Against Torture to use of autonomous robots for interrogation. • Overinclusive applicability of company law enabling incorporation of 'algorithmic entities' with corporate legal personhood.
	Obsolescence	Behaviour obsolete (necessity)	• New types of AI-supported remote biometric surveillance (gait or heartbeat identification) replace face recognition.
		Justifying assumptions no longer valid (adequacy)	• Structural unemployability through technological unemployment puts pressure on right to work, ILO regimes.
		No longer cost-effective (enforceability)	• Use of DeepFakes or computational propaganda raises monitoring and compliance enforcement costs for various regimes.
	Altered **problem portfolio** beyond institutional mandate/competency		• Military AI regime tailored to respond to ethical challenges of LAWS (e.g. maintaining meaningful human control over lethal force) might not be oriented to address risks of later adjacent AI capabilities (e.g. cyberwarfare) creating structural shifts.
Displacement	**Automation**	Law Creation & Adjudication	• Use of AI text-as-data tools to generate draft treaties, predict arbitral panel rulings, identify state practice, identify treaty conflicts.
		Monitoring & enforcement	• Improve depth & granularity of monitoring for treaty compliance • Increase breadth of monitoring by lowering participation threshold to other (e.g. non-state) actors • Improve actors' ability to make verifiable claims through architectural interventions
	Replacement	Changes in regulatory modality	• Use of AI tools such as emotion-recognition, social media sentiment analysis, or computational propaganda by states, resulting in increased state preference to resolve disputes in diplomatic channels.
Destruction	**Erosion** ('Development' intractable; gridlock)	Conceptual friction	• Attempted extension of existing regimes or norms to new technology cannot pass 'laugh test'.
		Political 'knots'	• Attempted extension of existing regimes or creation of new law, intractable because of political gridlock.
	Decline (increased contestation)	Increasing the spoils of noncompliance	• Innovations increase strategic stakes or ability to bypass monitoring, or lower proliferation thresholds or (political) noncompliance costs.
		Active weapon	• AI-enabled computational propaganda enables contestation of international law; • Suspected use of AI negotiation tools subverts legitimacy of resulting agreements.
		Shift of values	• AI capabilities perceived as enabling unilateralism, alternative to multilateralism

Regime Complexity: AI governance in 5 parts

	Theme	Questions
Origins Of individual regimes	**Purpose**: Is a regime needed?	• What are the underlying technological developments? • What (anticipated) **sociotechnical changes** do these enable? • What **governance rationales** are raised? (e.g. market failures; risks to human health; moral interests; social solidarity; democratic process, or international law itself) • What material features and **problem logics** characterize this governance target?
	Viability: (why) is any regime viable?	• From a comparative historical perspective, were past regimes for similar (technological) challenges viable? • Which (state) interests would this regime meet? What functions would it serve? • How might various actors shift norms to render it (more) viable?
	Design: what regimes optimal, adequate?	• What strategy? (e.g. reliance on (1) deterrence or (2) gradual norm development; (3) extension of regimes; (4) new regime) • If new regime, which **type**? (full ban or regulatory treaty?) Given differential resilience to **governance disruption**?
Topology of regime complex at a given time	**Demographics**	• **Size and composition** of network: what are the applicable norms or treaties, active institutions or governance initiatives?
	Organisation of network	• **Density** of institutional network (number of membership overlaps; institutional contact points on AI issue area) • **Type of links**: relating to norms, goals, impacts or institutional relations.
	Interactions and outcomes of linkages	• **Gaps**: functional non-regime, so issue unaddressed • **Conflictive links**: active norm conflicts, operational externalities, turf wars • **Cooperative links**: loose integration, but norm relationships unclear • **Synergistic links**: mutually reinforcing norms or institutional labour divisions
	Scope of analysis	• **Macro**: interactions of AI regime complex with other regimes (e.g. trade; data privacy; transport); or with general international law. • **Meso**: interactions of AI security regime with other AI regimes • **Micro**: internal institutional dynamics in AI security regime complex
Evolution given…	**General trends** in regime complexity?	• Density; accretion; power shifts over time; preference changes; modernity; representation and voice goals; local governance
	Effects of **AI governance disruption**?	• **Development**: AI as generator or trigger of latent regime fault lines • **Displacement**: AI as shield, patch, cure or accelerator of fragmentation. • **Destruction**: AI as driver of governance contestation
Consequences of trajectories…	If regime complex remains **fragmented**	• **Drawbacks**: undercuts coherence of international law; operational dysfunction; barriers to access and power inequalities; strategic vulnerability to forum shopping
		• **Benefits**: problem-solving; more democratic, inclusive; greater trust
	If regime complex is **integrated**	• **Drawbacks**: slowness, brittleness, 'breadth vs. depth' dilemma
		• **Benefits**: greater political power, efficiency and participation, can avert forum shopping
Strategies for managing AI regimes to ensure…	**Efficacy** (sociotechnical change)	• Conceptual approach (x3), instrument choice (x3), instrument design (x1)
	Resilience (governance disruption)	
	Coherence (regime complexity)	

Strategies

	Strategies for efficacy Sociotechnical change	**Strategies for resilience** Governance disruption	**Strategies for coherence** Regime complexity
Conceptual Approach	• Govern sociotechnical change, not technology • Triage governance rationales • Don't attempt to predict or wait; anticipate & adapt	• Expect 'Normal Disruption' of the global coordination architecture • Beware unreflexive technology analogies in treaty (re)interpetation • Pick your battles, beware legal hard-ball • Contain Digital Sovereignty and AI nationalism	• Consider AI issues in broader governance ecology • Consider avenues to shape regime foundations (interest, norms)

Instrument Choice	• New AI-application-specific regimes might be too siloed
	• Extending existing regimes to AI requires harmonisation
	• A global AI treaty might mistake AI's governance rationales
		• Treaties may be brittle; full bans could be resilient, but may not hold the door to AI disruption
		• Customary International Law as fall-back strategy
		• Standards over rules
		• Beware the unrestricted automation of international law—but recognize and promote cooperation-supportive AI tools
			• Choice between centralisation and decentralisation depends on trade-offs
			  • **Pro-centralization**: if AI governance depends more on political power, efficiency and accessible participation, ability to avert forum shopping
			  • **Pro-de-centralization:** if AI governance depends more on speed, adaptation, avoiding 'breadth-vs.-depth-dilemma'...
			• Explore adaptive instruments or strategies that mitigate or bypass trade-offs

Instrument Design	• Technology-neutral regulation foregrounded (if governance rationale is tech-neutral)	• Technology-neutral regulation • Pursue more flexible treaty designs (framework conventions; modular treaties; ...) • Let the future decide (e.g. authoritative interpreters)	• If a fragmented AI regime complex, foster regime interplay management / orchestration • If a centralised AI institution, design features for inclusion and adaptation

Further reading recommendations (selected)

Global Governance of AI

- Dafoe, Allan. "AI Governance: A Research Agenda." Oxford: Center for the Governance of AI, Future of Humanity Institute, 2018. https://www.fhi.ox.ac.uk/govaiagenda/.
- Cihon, Peter, Matthijs M. Maas, and Luke Kemp. "Fragmentation and the Future: Investigating Architectures for International AI Governance." *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.
- Kunz, Martina, and Seán Ó hÉigeartaigh. "Artificial Intelligence and Robotization." In *Oxford Handbook on the International Law of Global Security*, edited by Robin Geiss and Nils Melzer. Oxford University Press, 2021. https://papers.ssrn.com/abstract=3310421.
- Garcia, Eugenio V. "Multilateralism and Artificial Intelligence: What Role for the United Nations?" In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 18. Boca Raton: CRC Press, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3779866.
- Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. "Policy Brief: The Creation of a G20 Coordinating Committee for the Governance of Artificial Intelligence." *AI and Ethics*, October 6, 2020. https://doi.org/10.1007/s43681-020-00019-y.
- Schuett, Jonas. "A Legal Definition of AI." *ArXiv:1909.01095 [Cs]*, August 26, 2019. http://arxiv.org/abs/1909.01095.

AI and Automation in International Law

- Burri, Thomas. "International Law and Artificial Intelligence." *German Yearbook of International Law* 60 (October 27, 2017): 91–108. http://dx.doi.org/10.2139/ssrn.3060191
- Deeks, Ashley. "High-Tech International Law." *George Washington Law Review* 88 (2020): 575–653. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531976
- Boutin, Berenice. "Technologies for International Law & International Law for Technologies." *Groningen Journal of International Law* (blog), October 22, 2018. https://grojil.org/2018/10/22/technologies-for-international-law-international-law-for-technologies/.
- Kunz, Martina, and Seán Ó hÉigeartaigh. "Artificial Intelligence and Robotization." In *Oxford Handbook on the International Law of Global Security*, edited by Robin Geiss and Nils Melzer. Oxford University Press, 2020. https://papers.ssrn.com/abstract=3310421.
- Maas, Matthijs M. "AI, Governance Displacement, and the (De)Fragmentation of International Law." *ISA Annual Convention*, 2021. https://papers.ssrn.com/abstract=3806624.
- Dafoe, Allan, et al. "Open Problems in Cooperative AI." *ArXiv:2012.08630 [Cs]*, December 15, 2020. http://arxiv.org/abs/2012.08630.

Further reading recommendations (selected)

Military AI and arms control

- Maas, Matthijs M. "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons." *Contemporary Security Policy* 40, no. 3 (February 6, 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464.
- Rosert, Elvira, and Frank Sauer. "How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies." *Contemporary Security Policy* 0, no. 0 (May 30, 2020): 1–26. https://doi.org/10.1080/13523260.2020.1771508.
- Maas, Matthijs M. "Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot." *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 129–57. https://doi.org/10.1163/18781527-01001006.
- Rosert, Elvira, and Frank Sauer. "Prohibiting Autonomous Weapons: Put Human Dignity First." *Global Policy* 10, no. 3 (2019): 370–75. https://doi.org/10.1111/1758-5899.12691.
- Coe, Andrew J., and Jane Vaynman. "Why Arms Control Is So Rare." *American Political Science Review* 114, no. 2 (May 2020): 342–55. https://doi.org/10.1017/S000305541900073X.
- Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *ArXiv:2004.07213 [Cs]*, April 15, 2020. http://arxiv.org/abs/2004.07213. (Appendix)

Governance Disruption & (AI) Tech

- Crootof, Rebecca, and B. J. Ard. "Structuring Techlaw." *Harvard Journal of Law & Technology* 34 (forthcoming 2021). https://papers.ssrn.com/abstract=3664124.
- Liu, Hin-Yan, Matthijs Maas, John Danaher, Luisa Scarcella, Michaela Lexer, and Leonard Van Rompaey. "Artificial Intelligence and Legal Disruption: A New Model for Analysis." *Law, Innovation and Technology* 12, no. 2 (September 16, 2020): 205–58. https://doi.org/10.1080/17579961.2020.1815402.
- Crootof, Rebecca. "Jurisprudential Space Junk: Treaties and New Technologies." In *Resolving Conflicts in the Law*, edited by Chiara Giorgetti and Natalie Klein, 106–29, 2019. https://brill.com/view/book/edcoll/9789004316539/BP000015.xml.
- Crootof, Rebecca. "Regulating New Weapons Technology." In *The Impact of Emerging Technologies on the Law of Armed Conflict*, edited by Eric Talbot Jensen and Ronald T.P. Alcala, 1–25. Oxford University Press, 2019.
- Picker, Colin B. "A View from 40,000 Feet: International Law and the Invisible Hand of Technology." *Cardozo Law Review* 23 (2001): 151–219. https://papers.ssrn.com/abstract=987524
- Smith, Bryant Walker. "New Technologies and Old Treaties." *AJIL Unbound* 114 (ed 2020): 152–57. https://doi.org/10.1017/aju.2020.28.

Further reading recommendations (selected)

Regime Complexity & Architectures

- Morin, Jean-Frédéric, et al. "How Informality Can Address Emerging Issues: Making the Most of the G7." *Global Policy* 10, no. 2 (May 2019): 267–73. https://doi.org/10.1111/1758-5899.12668.
- Gómez-Mera, Laura, Jean-Frédéric Morin, and Thijs Van De Graaf. "Regime Complexes." In *Architectures of Earth System Governance: Institutional Complexity and Structural Transformation*, edited by Frank Biermann and Rakhyun E. Kim, 137–57. Cambridge University Press, 2020.
- Alter, Karen J., and Kal Raustiala. "The Rise of International Regime Complexity." *Annual Review of Law and Social Science* 14, no. 1 (2018): 329–49. https://doi.org/10.1146/annurev-lawsocsci-101317-030830.
- Biermann, Frank, Philipp Pattberg, Harro van Asselt, and Fariborz Zelli. "The Fragmentation of Global Governance Architectures: A Framework for Analysis." *Global Environmental Politics* 9, no. 4 (October 14, 2009): 14–40. https://doi.org/10.1162/glep.2009.9.4.14.

Legal Prioritization and Long-term Gov. Strategy

- Winter, Christoph, Jonas Schuett, Eric Martínez, Suzanne Van Arsdale, Renan Araújo, Nick Hollman, Jeff Sebo, Andrew Stawasz, Cullen O'Keefe, and Giuliana Rotola. "Legal Priorities Research: A Research Agenda." Legal Priorities Project, January 2021. https://www.legalpriorities.org/research_agenda.pdf.
- Liu, Hin-Yan, and Matthijs M. Maas. "'Solving for X?' Towards a Problem-Finding Framework to Ground Long-Term Governance Strategies for Artificial Intelligence." *Futures* 126 (February 1, 2021): 22. https://doi.org/10.1016/j.futures.2020.102672.
- Deudney, Daniel. "Turbo Change: Accelerating Technological Disruption, Planetary Geopolitics, and Architectonic Metaphors." *International Studies Review* 20, no. 2 (June 1, 2018): 223–31. https://doi.org/10.1093/isr/viy033.
- Stix, Charlotte, and Matthijs M. Maas. "Bridging the Gap: The Case for an 'Incompletely Theorized Agreement' on AI Policy." *AI and Ethics*, January 15, 2021. https://doi.org/10.1007/s43681-020-00037-w.